http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 6 No. 1, June 2017: 39 - 55
e-ISSN: 2289-2192

SEMANTIC MEASURE BASED ON FEATURES IN LEXICAL KNOWLEDGE SOURCES

UMMI ZAKIAH ZAINODIN
NAZLIA OMAR
ABDULGABBAR SAIF

ABSTRACT

Semantic measures between concepts require some of cognitive capabilities such as categorization and reasoning to estimate semantic association among concepts. For this reason, this problem has numerous applications in artificial intelligence, natural language processing, information retrieval, text clustering, and text categorization. Measuring lexical semantic relatedness generally requires certain background information about the concept or terms. Semantic measures between concepts are divided into two main sources: knowledge based and unstructured corpora. Both resources play important role in the task of measuring lexical semantic relatedness. Knowledge-based semantic measures have been proposed to estimate semantic similarity between two concepts using several approaches such as ontology-based, graph-based and concept's vector approaches. This paper reviews existing semantic similarity measures which depend on the lexical source and discusses the various approaches on semantic measures which include the path-based, information content, gloss-based and feature-based measures. This paper also focuses on semantic measures that are based on features using lexical knowledge sources and discusses some issues that arise in these measures.
Keyword: Semantic Measure, Knowledge Based, Semantic Similarity, Features-Based

PENGUKURAN SEMANTIK BERASASKAN CIRI DALAM PENGETAHUAN SUMBER LEKSIKAL

ABSTRAK

Pengukuran semantik antara konsep memerlu keupayaan kognitif, seperti pengkategorian dan alasan bagi mentafsir perkaitan semantik terhadap konsep. Masalah tersebut mempunyai aplikasi dalam kecerdasan buatan, pemprosesan bahasa tabii, capaian maklumat, pengelompokan teks, dan pengkategorian teks. Pengukuran hubungan semantik leksikal memerlu latar belakang maklumat tentang konsep atau terma. Pengukuran semantik antara konsep dibahagi kepada dua asas utama iaitu berasaskan pengetahuan dan berasaskan korpus tidak tersusun. Kedua-dua sumber ini memain peranan penting dalam tugasan pengukuran hubungan semantik leksikal. Pengukuran semantik berdasarkan sumber pengetahuan diperkenal bagi menganggar persamaan semantik di antara dua konsep mengguna beberapa pendekatan seperti berdasarkan ontologi, berdasarkan graf dan pendekatan konsep vektor. Kertas ini mengulas pengukuran persamaan semantik sedia ada yang bergantung kepada sumber leksikal dan membincang pelbagai pendekatan terhadap pengukuran semantik termasuk pengukuran berasaskan laluan, kandungan maklumat, berasaskan glos dan berasaskan ciri. Kertas ini turut memfokus terhadap pengukuran semantik berasaskan ciri dengan mengguna sumber pengetahuan leksikal dan membincang isu yang timbul berkaitan dengan pengukuran tersebut.
Kata Kunci: Pengukuran Semantik, Berasaskan Pengetahuan, Persamaan Semantik, Berasaskan Ciri

PENGENALAN

Perwakilan semantik merupakan topik popular dalam psikologi kognitif dan linguistik pengkomputeran yang berkait dengan bagaimana persamaan semantik dapat diwakilkan. Terdapat dua jenis perwakilan iaitu perwakilan semantik unit teks seperti perkataan, frasa atau dokumen berdasarkan sumber bahasa, dan perwakilan semantik sumber bahasa. Sumber bahasa

dapat diklasifikasi ke dalam sumber tidak berstruktur dan berstruktur. Bagi sumber tidak berstruktur (pengumpulan dokumen), terdapat beberapa kaedah yang dicadang mewakili semantik unit teks daripada maklumat yang bergantung kepada pendekatan statistik (Turney & Pantel, 2010). Bagi sumber berstruktur (sumber pengetahuan leksikal), bukti semantik diekstrak dari ciri yang dikod dalam sumber pengetahuan seperti hubungan leksikal (Agirre et al., 2009) dalam WordNet, atau hiperrangkai (Milne & Witten, 2008), kategori (Liberman & Markovitch ,2009), dan huraian teks konsep (Gabrilovich & Markovitch, 2007) dalam Wikipedia.

Perwakilan semantik berdasarkan sumber pengetahuan leksikal bergantung kepada pemetaan perkataan dalam teks bahasa tabii ke dalam unit (konsep) sepadan dalam sumber pengetahuan. Pengetahuan semantik boleh dikelola terutamanya mengguna dua model berlainan (Griffiths, Steyvers & Tenenbaum, 2007) iaitu model ruang semantik dan model rangkaian semantik. Model ruang semantik ialah satu model (Zesch & Gurevych, 2010) berasaskan ciri mewakili makna perkataan sebagai vektor atas konsep yang jelas berasal dari satu sumber pengetahuan berstruktur, selain daripada mengguna konsep yang tidak jelas daripada kebergantungan maklumat. Dalam rangkaian semantik model, perkataan dan konsep diwakili sebagai nod yang berhubungan antara satu dengan yang lain mengguna hubungan leksikal. Hubungan hipernim '*is a*' diguna bagi mewakili pengetahuan dalam sumber leksikal sebagai taksonomi semantik. Taksonomi semantik ialah satu rangkaian antara konsep dalam leksikon, nod mewakili konsep dan tepi mewakili hubungan hipernim.

Persamaan semantik merujuk kepada pengukuran yang bergantung ke atas sinonim dan hubungan '*is a*' dalam penilaian kekuatan semantik di antara dua konsep. Selain daripada itu, pengukuran perkaitan semantik merangkumi pelbagai hubungan antara konsep sama ada hubungan klasikal atau hubungan bukan klasikal. Tambahan pula, istilah jarak semantik merujuk kepada ukuran songsangan terhadap hubungan semantik atau persamaan.

Kertas ini mengulas pengukuran semantik berdasarkan sumber pengetahuan leksikal. Beberapa pengukuran semantik berdasarkan sumber pengetahuan leksikal yang diguna dalam pengukuran persamaan semantik berasaskan ciri turut dibincang selain daripada isu dalam pengukuran tersebut. Cadangan bagi penyelesaian isu tersebut turut dinyatakan.

KESAMAAN SEMANTIK PERKATAAN

Patwardhan, Banerjee dan Pedersen (2003) menyatakan kesamaan semantik di antara dua perkataan didefinisi sebagai hubungan di antara dua perkataan yang mentafsir satu persamaan. Pengiraan bagi kesamaan semantik di antara dua perkataan memberi satu nilai numerik. Keperluan data lingustik dalam pengiraan kesamaan semantik di antara dua perkataan memberi nilai jarak dalam kesamaan makna di antara dua perkataan, kewujudan kekerapan perkataan dalam koleksi serta jumlah persamaan sinonim antara perkataan tersebut. Berdasarkan nilai tersebut, ia dapat diguna dalam formula atau kaedah algoritma bagi menentu persamaan semantik di antara dua perkataan. Jadual 1 menunjukkan nilai persamaan semantik di antara dua perkataan yang mengguna formula kesamaan perkataan secara asimetri berasaskan korpus (Azmi-Murad & Martin, 2006).

JADUAL 1. Kesamaan semantik perkataan

$Perkataan_1$	$Perkataan_2$	Nilai semantik di antara dua perkataan
Province (Wilayah)	*Bay* (Teluk)	0.5
Watch (Pengawasan)	*Warn* (Amaran)	0.11
Evacuate (Berpindah)	*Homeless* (Tidak ada tempat tinggal)	1

Mihalcea, Corley dan Strapparava (2006) menyatakan kesamaan di antara dua perkataan dapat dibahagi kepada dua bahagian utama iaitu kesamaan berdasarkan korpus dan kesamaan berdasarkan pengetahuan. Kesamaan perkataan berasaskan korpus adalah persamaan semantik bagi mencari serta mengenal pasti kewujudan antara perkataan dengan mengguna data dari maklumat korpus yang besar seperti Korpus Brown, Korpus British Nasional, dan korpus Kolhapur. Antara kaedah atau algoritma yang diguna bagi mengira persamaan di antara dua perkataan berasaskan korpus ialah *Latent Semantic Analysis* (LSA), *Latent Dirichlet Allocation* (LDA) dan *Pointwise Mutual Information* (PMI). Bagi persamaan semantik di antara dua perkataan berasaskan pengetahuan adalah persamaan semantik yang mengira darjah hubungan antara perkataan mengguna sumber data dari kamus elektronik atau tesaurus (Resnik 1995). Antara contoh sumber leksikal data adalah seperti WordNet, *Longman Dictionary* dan tesaurus Roget. Bagi kaedah atau algoritma yang biasa diguna bagi mengira persamaan semantik antara perkataan berasaskan pengetahuan adalah seperti kaedah panjang laluan; kaedah Lecock dan Chodorow; kaedah Wu and Palmer dan kaedah Lesk.

TAKSONOMI SEMANTIK

Pengukuran hubungan semantik leksikal pada umumnya memerlu maklumat latar belakang tertentu tentang konsep atau istilah. Maklumat tersebut selalunya dikod dalam asas pengetahuan berstruktur dan separa berstruktur yang membentuk satu graf konsep yang berbentuk leksikal dan diindeks oleh bentuk kata. Konsep yang dihubung oleh pautan atau tepi menunjukkan satu maksud tertentu hubungan semantik. Bergantung kepada semantik pautan dan bentuk graf, istilah taksonomi merujuk kepada satu struktur hierarki dengan nod dikelola oleh hubungan pengkhususan generalisasi. Istilah ontologi merujuk kepada satu struktur taksonomi yang dirangkumi dengan hubungan semantik lain seperti antonim dan sinonim, ciri kelas atau sifat. Istilah semantik graf atau rangkaian semantik pula merujuk kepada jenis konsep graf yang dihubung oleh sebarang hubungan semantik atau hubungan pertautan longgar. Berdasarkan idea ini, taksonomi dapat didefinisi sebagai satu pengkhususan daripada ontologi, dan kedua-keduanya ialah jenis khusus graf semantik atau rangkaian. Semantik graf juga mengandungi pelbagai taksonomi atau ontologi. Jadual 2 menerangkan topologi parameter dalam ukuran taksonomi semantik.

JADUAL 2 Topologi parameter dalam ukuran taksonomi semantik

Parameter	Huraian keterangan
***path*(c$_1$, c$_2$)**	Panjang laluan terdekat di antara konsep $c_1$dan c_2 yang berkaitan dengan taksonomi
***descendants*(c)**	Set hiponim atau keturunan sesuatu konsep
***depth*(c)**	Kedalaman konsep *c* dalam hierarki taksonomikal
***ancestor*(c)**	Set yang dibentuk oleh semua pewaris sesuatu konsep *c*
***directParents*(c)**	Set induk bagi konsep c. Set ini termasuk dalam *ancestor*(c). Satu fungsi yang mengembali nod induk bagi sesuatu nod. Contohnya, *parent*(*Kereta api*) ={*kenderaan awam*}, dan *parent*(*Bata*)={*seramik*}. Kebanyakan kajian lepas mengguna 'induk' bagi merujuk kepada sebarang nod yang memasukkan nod dalam satu struktur taksonomi dengan *parent*(*c*) merujuk kepada pengumpulan nod dengan mengikuti pautan taksonomi daripada *c* kepada akar.
***leaves*(c)**	Set *leaves* yang berkaitan dengan *descendants*(c). Satu fungsi yang merujuk kepada nod anak sesuatu nod secara rekursi. Contohnya, *descendant* (*Kenderaan*)={*Kenderaan awam, Kereta api, bas*}, dan *descendant* (*Peralatan*)={*Kenderaan, Kenderaan awam, Kereta api, bas, Seramik, Bata*}.
***LCS*(c_1, c_2)**	*lowest common subsume* bagi konsep $c_1$dan c_2. Terdapat definisi lain bagi *lcs* dalam susastera. Majoriti susastera mengguna definisi yang ditakrif oleh Resnik (1995), yang mendefinisi *lcs* sebagai ahli dalam *cs*(c_1, c_2) pada paras terendah bagi taksonomi. Maka, *lcs*(*Kereta api, Bas*)={*Kenderaan awam*}, dan *lcs* (*Bas, Bata*)={*Peralatan*}.

nod	Nod yang merujuk kepada konsep dalam graf semantik. Mengguna nod dan konsep kesalingbolehtukaran apabila menggambar taksonomi dan graf semantik.
Tepi (*edge*)	Satu pautan yang menghubung dua nod bersebelahan dalam satu graf semantik. Satu graf semantik menghubung nod bagi hubungan tertentu, dengan menyatakan ciri sesuatu tepi dapat ditakrif oleh hubungan yang diwakili. Jarak tepi antara "Kenderaan" dan "Seramik" ialah 2. Jumlah tepi dalam contoh taksonomi tersebut mewakili hubungan "adalah satu" di antara dua nod.
Akar (*root*)	Akar nod dalam T, dalam kes ini, nod entiti.
***subsumer*(*c*)**	Satu fungsi yang merujuk pada *subsumers* sesuatu nod secara rekursi. Contoh: *subsumer*(*Objek*)= {*Entiti fizikal, entiti*}, dan *subsumer*(Bata) = {*Seramik,Peralatan, Artifak,Objek,Entiti Fizikal, Entiti*}. '*Subsumer*' juga diguna bagi kesalingbolehtukaran dengan 'pewaris', atau hipernim.
child(c)	Satu fungsi yang merujuk kepada nod anak sesuatu nod. Contoh: *child* (*Peralatan*)={*Kenderaan, Seramik*}.

SUMBER PENGETAHUAN LEKSIKAL

Sumber pengetahuan leksikal penting untuk tugas pemprosesan bahasa tabii seperti, pengukuran persamaan semantik, pengecaman entiti nama, dan penyahtaksaan makna perkataan. Sumber pengetahuan leksikal secara elektronik dibangun oleh ahli bahasa seperti, *Longman Dictionary* oleh (Paul 1978), tesaurus Roget oleh (Roget, 1911), dan WordNet oleh (Fellbaum, 1998). Terdapat beberapa sumber pengetahuan pelbagai bahasa yang dicipta dan diselenggara berdasarkan usaha bersama seperti Wiktionary and Wikipedia selain daripada beberapa sumber pengetahuan dalam domain bioperubatan yang diguna bagi menilai tugas persamaan semantik seperti *Systematized Nomenclature of Medicine-Clinical Terminology* (SNOMED CT); *Medical Subject Headings* (MeSH); dan *UMLS Metathesaurus*.

***WordNet*:** WordNet adalah sebuah kamus yang boleh dipadan dengan teori ingatan semantik manusia. WordNet (WN) oleh Fellbaum (1998) ialah pangkalan data leksikal besar berdasarkan prinsip psikolinguistik bagi bahasa Inggeris yang mengandungi set sinonim (*synsets*) dan hubungan (hubungan leksikal antara *synsets*). Setiap *synsets* mewakili konsep tunggal yang dikenal pasti oleh set istilah lema (*lemma*). *Synsets* dibahagi berdasarkan bahagian ucapan (POS) ke dalam empat jenis iaitu kata nama, kata kerja, kata sifat, dan adverba.

WordNet menyedia dua ciri semantik utama yang boleh diguna dalam bidang linguistik pengkomputeran iaitu glos dan hubungan semantik. Glos ialah satu bahagian yang penting dalam WordNet dan mengandungi takrif teks pendek atau pengagihan *synsets*. *Synsets* dalam WordNet dikaitkan antara satu dengan yang lain melalui pelbagai jenis hubungan semantik dan leksikal yang boleh diguna sebagai ciri untuk mengukur secara langsung persamaan semantik konsep atau *synsets*. Setiap *synsets* mempunyai frasa definisi, contoh atau komen tambahan.

PENGUKURAN PERSAMAAN SEMANTIK BERASASKAN PENGETAHUAN

Pengukuran semantik berdasarkan pengetahuan merujuk pada kaedah analisis semantik yang bergantung kepada pengetahuan yang dikod dalam sumber kelolaan manusia. Resnik (1995) mendefinisi pengukuran kesamaan semantik berdasarkan pengetahuan adalah kesamaan semantik yang mengira darjah hubungan antara perkataan dengan mengguna sumber data dari kamus elektronik dan tesaurus.

LANGKAH BERASASKAN LALUAN

Pengukuran berasaskan laluan bergantung kepada jarak antara nod dalam taksonomi semantik bagi menentu persamaan semantik di antara dua konsep. Rada et al. (1989) mengusul

pengukuran pertama jarak semantik adalah sebagai panjang laluan terdekat antara nod dalam graf semantik. Laluan terdekat panjang dua ruas $\delta(c_1, c_2)$ didefinisi sebagai bilangan minimum tepi yang memisah nod. Jarak semantik dinormalisasi untuk mengira persamaan semantik mengguna fungsi salingan, dan logaritma negatif jarak atas kedalaman berganda taksonomi (Leacock & Chodorow 1998) semantik seperti formula berikut:

$$Sim_{Rada}(c_1, c_2) = \frac{1}{\delta(c_1,c_2)+1} \quad (1)$$

$$Sim_{LCh}(c_1, c_2) = -\log\left(\frac{\delta(c_1,c_2)+1}{2\times \mathrm{D}}\right) \quad (2)$$

dengan $\delta(c_1, c_2)$ ialah laluan terdekat atau terpendek di antara c_1 dan c_2, dan D ialah kedalaman taksonomi semantik atau nilai maksimum sesuatu taksonomi.

Ukuran semantik hanya bergantung kepada jarak dan tidak mengambil kira konsep khusus dalam organisasi taksonomi semantik. Dua pasangan nod mempunyai jarak yang sama mempunyai persamaan semantik yang sama tanpa mengambil kira darjah (kedalaman) khusus dalam taksonomi semantik. Bagi menangani had panjang laluan terdekat, dua ciri iaitu jarak di antara nod dan kedalaman subsumer (*Least Common Subsumer*) nod tersebut digabung. Pengukuran kesamaan semantik didefinisi sebagai di antara dua ruas (c_1 dan c_2) dalam taksonomi semantik bagi menyepadu jarak dan kedalaman *Least Common Subsumer* (LCS) dalam contoh yang lain (Li, Bandar & Mclean, 2003; Liu, Zhou & Zheng, 2007; Wu & Palmer, 1994) seperti berikut:

$$Sim_{WuP}(c_1, c_2) = \frac{2\times depth(LCS(c_1,c_2))}{\delta(c_1,c_2)+2\times depth(LCS(c_1,c_2))} \quad (3)$$

$$Sim_{Li}(c_1, c_2) = e^{-\alpha\times\delta(c_1,c_2)} \times \frac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}}\ ; h = depth(LCS(c_1, c_2)) \quad (4)$$

$$Sim_{Liu}(c_1, c_2) = \frac{\beta\times depth(LCS(c_1,c_2))}{\beta\times depth(LCS(c_1,c_2))+\alpha\times\delta(c_1,c_2)} \quad (5)$$

dengan α and β adalah parameter yang menskala penyumbangan terhadap panjang laluan terdekat dan kedalaman LCS.

Taieb, Aouicha dan Hamadou (2014) mengusul pengukuran persamaan semantik yang bergantung pada dua topologi parameter berkaitan dengan 'is a' taksonomi: hiponim dan parameter kedalaman. Hiponim diguna bagi menentu spesifikasi konsep kepunyaan WordNet 'is a' taksonomi. Parameter kedalaman diguna bagi mengukur spesifikasi sesuatu konsep dalam perspektif lain kerana apabila sesuatu tahap meningkat dalam taksonomi dari satu tahap ke tahap yang lain, perambatan data ke arah keturunan dengan tambahan spesifikasi tertentu terhasil. Pengukuran semantik di antara dua konsep c_1 dan c_2 dikira seperti berikut:

$$Sim_{T14}(c_1, c_2) = |TermDepth(c_1, c_2)| \times TermHypo(c_1, c_2) \quad (11)$$

dengan $TermDepth(c_1, c_2)$ menggambar parameter kedalaman yang dikira mengguna koefisien Dice pada kedalaman seperti berikut:

$$TermDepth(c_1, c_2) = \frac{2\times Depth\ (LCS(c_1,c_2))}{Depth\ (c_1)+Depth\ (c_2)} \quad (12)$$

$TermHypo(c_1, c_2)$ merujuk kepada parameter hiponim yang dikira sebagai koefisien Dice termasuk LCS yang mewakili pekongsian maklumat antara konsep sasaran:

$$TermHypo(c_1, c_2) = \frac{2\times Spec_{Hypo}(LCS(c_1,\ c_2))}{Spec_{Hypo}(c_1)+Spec_{Hypo}(c_2)} \quad (13)$$

$Spec_{Hypo}(c)$ ialah spesifisiti konsep c yang diberi seperti berikut:

$$Spec_{Hypo}(c) = 1 - \frac{\log(HypoValue(c))}{\log(HypoValue(root))} \quad (14)$$

dengan $HypoValue(c)$ ialah kuantiti hiponim dalam taksonomi yang ditakrif bergantung kepada taburan kebarangkalian kedalaman atas taksonomi semantik sebagai:

$$HypoValue(c) = \sum_{c' \in Hypo(c)} P(depth(c')) \quad (15)$$

Hypo(c) ialah hiponim konsep c dan kedalaman (c) merujuk kepada panjang laluan jauh di antara c dan akar taksonomi. $P(depth(c'))$ ialah taburan kebarangkalian terhadap taksonomi semantik seperti berikut:

$$P\big(depth(c')\big) = \frac{|c' \in C | depth(c') = depth(c)|}{N} \quad (16)$$

C ialah konsep yang berkaitan dengan WordNet 'is a' taksonomi dan N ialah kardinaliti terhadap set *C*.

LANGKAH BERDASARKAN KANDUNGAN MAKLUMAT

Pengukuran berdasarkan kandungan maklumat diperkenal bagi permodelan persamaan semantik di antara dua konsep leksikal dalam taksonomi tertentu berdasarkan takat berkongsi maklumat persamaan. Konsep yang mempunyai nilai kandungan maklumat yang tinggi dianggap spesifik berbanding dengan konsep yang mempunyai nilai yang rendah. Pengukuran persamaan semantik di antara dua konsep (Resnik, 1995) dikira sebagai IC subsumer yang sepunya terkecil bagi konsep seperti berikut:

$$Sim_{Res}(c_1, c_2) = IC(LCS(c_1, c_2)) \quad (17)$$

Ukuran ini bergantung pada kandungan maklumat konsep subsumer sepunya terkecil yang membawa isu 'ketakbolehbezaan' (nilai persamaan semantik di antara pasangan konsep yang mempunyai LCS yang seiras) (Budanitsky & Hirst, 2006). Kajian (Jiang & Conrath, 1997; Lin, 1998; Pirró & Euzenat, 2010) menangani isu ini dengan penskalaan kandungan maklumat konsep LCS oleh kandungan maklumat konsep tersendiri. Terdapat dua langkah terkenal yang mentakrif persamaan semantik sebagai formula berikut:

$$Sim_{lin}(c_1, c_2) = \frac{2 \times Ic(LCS(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (18)$$

$$Sim_{PE}(c_1, c_2) = \frac{Ic(LCS(c_1, c_2))}{IC(c_1) + IC(c_2) - Ic(LCS(c_1, c_2))} \quad (19)$$

KANDUNGAN MAKLUMAT KOMPUTERAAN

Pengukuran semantik berasaskan kandungan maklumat komputeraan antara ukuran terawal diperkenal oleh (Resnik, 1995) bagi menggabung bukti semantik dari hierarki semantik dan maklumat korpus yang dilabel. Dalam ukuran ini, konsep kandungan maklumat ditentu sebagai negatif logaritma kebarangkalian *p(c)* berhadapan contoh konsep dalam satu korpus yang dilabel sebagai formula berikut:

$$IC_{Res}(c) = -\log p(c) \quad (20)$$

Pengukuran kandungan maklumat (*Information Content (IC)*) adalah konsep yang memerlu korpus yang besar dan cekap serta ontologi semantik. Bagaimanapun, korpus yang sedia ada dalam beberapa bahasa seperti Bahasa Melayu memberi cabaran utama bagi menilai pengukuran semantik. Bagi mengatasi masalah ini, beberapa kajian dilakukan bagi mentaksir kandungan maklumat dengan hanya mengguna struktur hierarki sumber pengetahuan leksikal. Kajian ini dipanggil berasaskan ontologi atau kaedah IC Intrinsik yang membawa keluasan makna atau spesifikasi konsep tertentu yang dapat dianggar berdasarkan ciri berlainan tentang kedudukan konsep dalam taksonomi semantik.

Berdasarkan taksonomi semantik, Seco, Veale dan Hayes (2004) memperkenal kaedah berasaskan ontologi pertama yang mengukur IC intrinsik konsep berdasarkan andaian dengan menyatakan organisasi ontologi adalah berstruktur dan bermakna maka konsep yang mempunyai banyak hiponim menghasil kurang maklumat daripada konsep daun. Pengiraan IC sesuatu konsep mengabai tahap keluasan konsep. Walaupun sesuatu konsep mempunyai tahap yang tinggi dalam ontologi taksonomi, tetapi maklumat dalam konsep tersebut adalah berkurangan. Kandungan maklumat konsep tertentu ditakrif bergantung pada jumlah hiponimnya seperti berikut:

$$IC_{Seco}(c) = 1 - \frac{\log(hypo(c)+1)}{\log(Max_{nodes})} \tag{21}$$

Hypo(c) ialah jumlah keturunan *c* dalam taksonomi, dan Max_{nodes} ialah jumlah keseluruhan konsep dalam taksonomi. Zhou, Wang dan Gu (2008) mengusul satu model bagi mengira IC dengan menyatu hiponim dan ciri kedalaman konsep tertentu seperti berikut:

$$IC_{Zhou}(c) = k\left(1 - \frac{\log(hypo(c)+1)}{\log(Max_{nodes})}\right) + (1-k)\left(\frac{\log(depth(c))}{\log(Max_{depth})}\right) \tag{22}$$

dengan Max_{depth} ialah kedalaman maksimum hierarki semantik, dan *k* ialah satu parameter penalaan yang mengimbangi penyumbangan hiponim dan ciri kedalaman.

Sánchez, Batet dan Isern (2011) menentu ukuran kandungan maklumat konsep berdasarkan daun dan *subsumers* konsep yang bersesuaian dengan darjah keluasan dan ketepatan dalam taksonomi semantik tersendiri. Ukuran ini dirumus sebagai formula berikut:

$$IC_{Sanchez}(c) = -\log\left(\frac{\frac{|\text{Leaves(c)}|}{|subsumers(c)|}+1)}{Max_{Leaves}+1}\right) \tag{23}$$

dengan Max_{Leaves} ialah jumlah daun dalam taksonomi, *subsumers(c)* merujuk kepada hipernim satu konsep c dengan cara rekursi, dan *Leaves* ialah set konsep pada penghujung pokok taksonomi di bawah konsep c. Mengguna ciri yang sama, Sánchez et al. (2012) mentakrif semula kandungan maklumat konsep dalam graf semantik dengan mengguna pengertian kebiasaan. Kebiasaan dan IC konsep dalam hierarki semantik dikira seperti berikut:

$$commonness(c) = \begin{cases} \dfrac{1}{|subsumers(c)|} & where\ c\ is\ a\ leaf \\ \displaystyle\sum_{c' \in \text{Leaves(c)}} \frac{1}{|subsumers(c')|} & otherwise \end{cases}$$

$$IC_{SanBatet}(c) = -\log\left(\frac{commonness(c)}{commonness(Root)}\right) \tag{24}$$

Taieb et al. (2011) memperkenal kaedah untuk mengira kandungan maklumat konsep dalam taksonomi semantik seperti berikut. Konsep yang diberi dilambang sebagai vektor *V* yang mengandungi konsep tersendiri dengan set pengumpulan pewaris dalam taksonomi berhierarki. Seterusnya bagi setiap unsur dalam vektor itu ditetap untuk skor pemberatan dari unsur tersebut ke dalam kandungan maklumat terhadap konsep yang diberi. Skor setiap konsep dikira mengguna tiga ciri iaitu induk langsung (hipernim), keturunan (hiponim), dan kedalaman seperti formula berikut:

$$Score(c) = \left(\sum_{a \in Hyper(c)} \frac{depth(a)}{Hypo(a)}\right) \times Hypo(c) \quad (25)$$

Hyper(c) ialah jumlah hiponim *c* dalam taksonomi. Kandungan maklumat konsep diukur dengan skor setiap konsep dalam vektor perwakilan konsep yang diberi seperti berikut:

$$IC_{Taieb}(c) = (\sum_{c' \in V} Score(c')) \times AvgDepth(c) \quad (26)$$

AvgDepth (c) ialah fungsi purata kedalaman konsep c dalam vektor perwakilan. Purata kedalaman menunjukkan maklumat tentang taburan vertikal konsep pewaris seperti berikut:

$$AvgDepth(c) = \frac{1}{|V|} \times \sum_{c' \in V} depth(c') \quad (27)$$

Meng, Gu dan Zhou (2012) mencadang model yang menggabung kedalaman dan hiponim bagi mengukur konsep IC dalam graf semantik. Model ini ditakrif seperti berikut:

$$IC_{Meng}(c) = \frac{\log(depth(c))}{\log(Max_{depth})} \times \left(1 - \frac{\log\left(\sum_{c' \in Hypo(c)} \frac{1}{depth(c')} + 1\right)}{\log(Max_{nodes})}\right) \quad (28)$$

Hypo(c) ialah jumlah hiponim konsep c, Max_{depth} ialah kedalaman maksimum dalam taksonomi semantik dan Max_{nodes} ialah jumlah nod dalam taksonomi semantik.

Ben Aouicha, Hadj Taieb dan Hamadou (2016) mengusul pengukuran baharu bagi mengira kandungan maklumat konsep berdasarkan lima parameter topologi iaitu kedalaman, pewaris, hipernim, keturunan dan hiponim. Untuk setiap konsep, IC ditakrif berdasarkan sub-graf oleh pewaris seperti berikut:

$$IC_{Aouicha}(c) = \sum_{c' \in Subgraph\ (c)} Score(c') \quad (29)$$

Skor konsep *c'* ditentu mengguna hiponim seperti berikut:

$$Score(c) = \left(\sum_{c' \in Hyper(C)} \left(-\log\left(\frac{1}{depth(c')}\right) \times Term(c')\right)\right) \times Term(c) \quad (30)$$

Term merujuk kepada setiap konsep kepunyaan sub-graf sesuatu konsep yang ditakrif seperti berikut:

$$Term(c) = 1 - \frac{\log(HypoInfo(c)+1)}{\log(maxHypoInfo)} \quad (31)$$

$HypoInfo(c)$ didefinisi sebagai:

$$HypoInfo(c) = \sum_{c' \in descendants(c)} \frac{1}{depth\ (c')} \quad (32)$$

UKURAN BERASASKAN TINDIHAN (*OVERLAP*)

Ukuran berasaskan tindihan bergantung kepada jumlah pengetahuan umum di antara dua konsep. Ukuran berasaskan tindihan terawal ialah teknik yang merangkumi glos atau penerangan dalam sumber leksikal. Penjelasan ringkas ini telah dieksploitasi bagi mengukur hubungan semantik di antara dua konsep dengan mencari pertindihan teks antara glos langsung (Lesk, 1986) atau glos lanjutan (Banerjee & Pedersen, 2003). Hubungan semantik di antara c_1 dan c_2 ditakrif mengikut skala pertindihan oleh panjang minimum (*min*) dua set seperti yang ditunjuk dalam formula berikut:

$$SRel_{Lesk}(c_1, c_2) = \frac{|g(c_1) \cap g(c_2)|}{\min(|g(c_1)|, |g(c_2)|)} \quad (33)$$

$|g(c_i)|$ adalah panjang glos konsep c_i.

Banerjee dan Pedersen (2003) mencadang skor berasaskan tindihan dengan menganggap ungkapan pemberat pelbagai perkataan dalam perwakilan teks melebihi daripada pemberat satu perkataan. Skor pertindihan ini ditakrif untuk *n* frasa *m-word* yang ada dalam dua buah teks seperti berikut:

$$BP_{overlap}(g(c_1), g(c_2)) = \sum_n m^2 \quad (34)$$

Bagi mengukur hubungan semantik, Strube dan Ponzetto (2006) menormalisasi ukuran berasaskan tindihan dengan mengambil kira saiz setiap glos yang diwakili diset kepada dua konsep seperti yang ditunjuk dalam formula berikut:

$$SRel_{SP}(c_1, c_2) = \tanh\left(\frac{BP_{overlap}(g(c_1), g(c_2))}{|g(c_1)| + |g(c_2)|}\right) \quad (35)$$

tanh (x) ialah fungsi tangen hiperbolaan.

Selain daripada itu, ukuran berasaskan tindihan bergantung kepada pengekstrakan pengetahuan dari taksonomi semantik sumber leksikal selain daripada glos dalam kajian sebelumnya. Gentleman (2005) memperkenal ukuran semantik berdasarkan pengetahuan yang dikongsi oleh pewaris di antara dua konsep yang diekstrak dari taksonomi semantik. Ukuran ini ditakrif seperti berikut:

$$Sim_{Gen}(c_1, c_2) = \frac{|ancestor(c_1) \cap \text{ancestor}(c_2)|}{|ancestor(c_1) \cup ancestor(c_2)|} \quad (36)$$

Batet, Sánchez dan Valls (2011) mengusul pengukuran semantik berasaskan tindihan bergantung kepada jumlah maklumat daripada konsep pewaris ke dalam taksonomi semantik. Dalam ukuran ini, konsep dilambang sebagai satu set super-konsep yang boleh dihasil dari *subsumers* satu konsep ke dalam semua laluan taksonomikal daripada konsep ke akar. Super-konsep set diekstrak bagi konsep khusus dalam graf semantik *C* seperti berikut:

$$T(c_i) = \{c_j \in C : c_j \; is \; a \; \text{superconcept of } c_i\} \cup \{c_i\} \quad (37)$$

Oleh yang demikian, persamaan semantik di antara dua konsep c_1 dan c_2 ditakrif sebagai nisbah antara jumlah konsep bukan berkongsi dan jumlah dikongsi dan konsep bukan berkongsi seperti berikut:

$$Sim(c_1, c_2) = -\log_2 \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \quad (38)$$

KAJIAN LAMPAU TENTANG PENGUKURAN SEMANTIK BERASASKAN CIRI

Pengukuran semantik berasaskan ciri adalah tidak bersandar pada taksonomi dan *subsumers* sesuatu konsep, selain daripada mengeksploitasi sifat ontologi untuk mendapat nilai persamaan. Hal yang sedemikian berdasarkan kepada andaian setiap konsep digambar oleh set perkataan yang merujuk kepada sifat atau ciri seperti definisi atau glos dalam WordNet. Semakin banyak ciri kebiasaan yang dipunyai oleh dua konsep, dan semakin kurang ciri bukan kebiasaan yang dipunyai oleh mereka, semakin sama persamaan konsep tersebut. Pengukuran klasik iaitu Tversky model, menganggap persamaan adalah tidak simetrik. Ciri antara subkelas dan superkelasnya mempunyai sumbangan yang besar kepada penilaian persamaan serta dapat didefinisi sebagai:

$$Sim_{Tversky}(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + k|C_1/C_2| + (k-1)|C_2/C_1|} \quad (39)$$

dengan c_1 dan c_2 sepadan dengan huraian set konsep c_1 dan c_2, *k* boleh dilaras dan k∈[0,1]. Dari formula tersebut, dapat dinyata nilai ukuran ini berbeza dari 0 hingga 1. Selain daripada itu, *sim* bertambah dengan persamaan dan berkurang dengan perbezaan di antara dua konsep.

Definisi set ciri adalah genting dalam model ini. Goodman (1972) mempertikai penaksiran persamaan di antara objek *a* dan *b* adalah kabur dan tidak bermakna tanpa satu "rangka rujukan". Pertanyaan soalan "Bagaimana persamaan a dan b?" memberi jawapan kepada soalan yang berbeza "Bagaimanakah a dan b adalah sama?" (Medin, Goldstone & Gentner, 1993). Pendekatan sedia ada bergantung kepada maklumat yang boleh didapati dalam ontologi, khususnya set sinonim (dipanggil *synsets* dalam WordNet Fellbaum (1998)), definisi (*glosses*, mengandungi huraian teks makna perkataan) dan pelbagai jenis hubungan semantik yang dipertimbang.

Rodríguez dan Egenhofer (2003) mengusul pendekatan mengira persamaan semantik. Persamaan dikira sebagai hasil tambahan wajaran persamaan antara *synsets*, ciri (meronim dan sifat) dan konsep jiran (dihubung melalui penunjuk semantik) seperti penilaian seperti berikut:

$$Sim(a, b) = w \times S_{synset}(a, b) + u \times S_{features}(a, b) + v \times S_{\text{neighborhoods}}(a, b) \quad (40)$$

dengan S_{synset}, $S_{features}$ dan $S_{neighborhoods}$ ialah persamaan antara set sinonim, ciri-ciri, dan terma penilaian kejiranan semantik, *w*, *u*, dan *v* adalah pemberat persamaan bagi setiap spesifikasi komponen yang bergantung kepada sifat ontologi dan *S* mewakili pertindihan antara perbezaan ciri yang dikira seperti berikut:

$$S(a, b) = \frac{|A \cap B|}{|A \cap B| + \alpha(a,b)|A \backslash B| + (1-\alpha(a,b))|B \backslash A|} \quad (41)$$

dengan *A*, *B* ialah terma yang dinilai untuk konsep sepadan bagi *a* dan *b*, A/B adalah set terma bagi *A* tetapi tiada dalam *B* (songsangan untuk B/A), dan |.| adalah kekardinaliti sesuatu set. *a* pula adalah satu fungsi yang menentu kepentingan relatif ciri bukan kebiasaan, secara formal *a(a,b)*, dikira sebagai fungsi kedalaman a dan b dalam taksonomi seperti berikut:

$$\alpha(a, b) = \begin{cases} \left(\frac{depth(a)}{depth(a)+depth(b)}\right), depth(a) \leq depth(b) \\ \left(1 - \frac{depth(a)}{depth(a)+depth(b)}\right), depth(a) > depth(b) \end{cases} \quad (42)$$

Kedalaman (*a*) merujuk kepada laluan terdekat dari terma penilaian *a* kepada bayangan akar, dan kedalaman ini mencermin darjah granualiti yang mereka bentuk ontologi tersebut.

Fungsi berdasarkan ciri mengganggap persamaan-X bergantung ke atas kesepadanan antara *synsets* dan set huraian istilah (Petrakis et al. 2006). Set huraian istilah mengandungi

perkataan yang diekstrak oleh definisi istilah penghuraian (*glosses*) dalam WordNet (Fellbaum, 1998) atau 'nota skop' dalam MeSH (Sánchez & Batet, 2013). Dua terma adalah sama jika *synsets* atau set huraian atau *synsets* terma dalam kejiranan (lebih khusus dan lebih banyak terma umum) adalah sama dari segi leksikal. Fungsi persamaan dinyata seperti berikut:

$$Sim_x(a,b) = \begin{cases} (1\ ), & if\ S_{synset}(a,b) > 0 \\ \left(\max\left(S_{description}(a,b), S_{neighborhoods}(a,b)\right)\right), & if\ S_{synset}(a,b) = 0 \end{cases} \quad (43)$$

Persamaan untuk jiran semantik $S_{\text{neighborhoods}}$ dikira seperti berikut:

$$S_{neighborhoods}(a,b) = \max\left\{\frac{|A_i \cap B_i|}{|A_i \cup B_i|}\right\} \quad (44)$$

dengan *i* menanda jenis hubungan. Oleh kerana tidak semua terma dalam kawasan kejiranan terma dihubung dengan hubungan yang sama, maka setiap hubungan semantik yang berbeza dikira secara maksimum dan berasingan. Persamaan untuk huraian set $S_{description}$ dan sinonim $S_{synsets}$ dikira seperti berikut:

$$S(a,b) = \frac{|A \cap B|}{|A \cup B|} \quad (45)$$

dengan A and B merujuk kepada set *synsets* atau huraian terma bagi terma *a* dan *b*.

Jiang et al. (2015) mengusul pengukuran semantik berasaskan ciri mengguna struktur Wikipedia sebagai ciri semantik. Setiap artikel Wikipedia diandai dengan menggambar satu entiti lengkap, yang dipanggil konsep (satu gambaran lengkap maklumat dilukis oleh satu artikel Wikipedia). Konsep ditentu oleh empat ciri iaitu nama pengecam, set alias alternatif atau sinonim, glos Wikipedia, set teks penambat beberapa konsep Wikipedia dan kategori. Tajuk artikal Wikipedia ialah frasa ringkas dalam sebuah tesaurus konvensional. Perwakilan formal konsep Wikipedia *Con* ditakrif seperti berikut:

$$Con = \{\text{synonyms}, \text{glosses}, \text{anchors}, \text{categories}\}$$

Dengan Synonyms = $\{Con_1, \ldots, Con_m\}$ adalah set sinonim *Con*, *Glosses* ialah perenggan pertama teks A, *Anchors* = $\{Anc_1, \ldots, Anc_n\}$ adalah set teks penambat (label hiperrangkai dalaman) dalam A, dan *Categories* = $\{Cat_1, \ldots, Cat_k\}$ adalah set kategori A.

Bagi mewakili semantik nod, Saif, Ab Aziz dan Omar (2017) mengguna taksonomi semantik WordNet(WN) berdasarkan hubungan berhierarki dalam kalangan *synsets*. Untuk setiap nod (*synset*), perwakilan dihasil dengan merangkai lema (*lemma*) dalam konsep yang berkaitan (pewaris, anak, dan adik-beradik) sebagai teks (jujukan perkataan). Pewaris bagi *synset* diekstrak sebagai nod laluan yang merujuk kepada *synset* tersebut kepada akar taksonomi. Nod pada pelbagai tahap dalam taksonomi semantik mempunyai perbezaan kemungkinan bilangan pewaris, nod dalam paras terendah mempunyai banyak pewaris daripada nod dalam paras tertinggi. Nod boleh mempunyai lebih daripada satu laluan kepada akar (pelbagai hubungan hipernim), maka pewaris diekstrak dari setiap laluan.

Jumlah pewaris bagi nod ditentu berdasarkan kedalaman nod tersebut dalam hierarki semantik. Untuk satu nod *v* dengan laluan kepada akar, jumlah pewaris terdekat dari setiap laluan *p* yang sepatutnya dimasukkan dalam glos ditakrif oleh formula berikut:

$$N_{\text{ancestors}}(p_v) = \text{Round}\left(\frac{\text{Length}(p_v) \times 4}{Max_{depth}}\right) \quad (46)$$

Max_{depth} ialah kedalaman taksonomi WN, *Length* (p_v) menyatakan panjang laluan terma p_v dalam jumlah tepi, dan *Round* ialah satu fungsi pusingan angka *x* kepada integer (*Round(1.5) =2, Round(1.3)=1*). Bagi setiap laluan daripada setiap nod kepada akar, formula ini menyatakan satu angka berjarak di antara 0 dan 4.

Anak *synset* adalah nod yang secara terus memasukkan (hubungan hiponim) oleh *synset* dalam taksonomi. Oleh kerana kebanyakkan nod adalah daun (*leaves*) dalam taksonomi WN (mereka tidak mempunyai anak), adik-beradik setiap nod diekstrak sebagai konsep berkaitan selain daripada anak. Adik-beradik *synset* ialah nod yang mempunyai induk yang sama dalam *synset*. *Synset* boleh menjadi perbezaan bilangan anak atau adik-beradik merentasi nod dalam taksonomi. Maka, setiap nod diwakili oleh maksimum empat nod daripada anak atau adik-beradiknya. Untuk *synset* yang mempunyai lebih daripada empat nod berkaitan, berdasarkan spesifikasi nod dalam taksonomi, empat nod teratas daripada anak atau adik-beradiknya dipilih sebagai konsep berkaitan *synset* tersebut. Spesifikasi nod v ditakrif sebagai panjang laluan terdekat antara nod dan daun-daunnya oleh formula berikut:

$$specificity(v) = \frac{1}{\min\limits_{c \in leaves(v)} \{\text{length}(v,c)\}} \tag{47}$$

dengan fungsi *length* merujuk kepada jumlah nod dalam laluan terdekat antara dua *synsets* dalam taksonomi.

Menurut (Gabrilovich & Markovitch, 2007) kaedah Analisis Semantik Eksplisit (ESA) diperkenal sebagai vektor model ruang bagi mentafsir perkataan semantik berdasarkan pengetahuan yang dikod dalam artikel Wikipedia. Idea utama di sebalik kaedah ESA ini ialah makna perkataan boleh dinyata dengan konsep yang tak tersirat dalam sumber bahasa. Model klasik bagi vektor ruang diguna bagi membina matriks perkataan-konsep bergantung ke atas taburan perkataan dalam artikel Wikipedia. Dalam matriks tersebut, setiap unsur dikira mengguna teknik TF-IDF. Interpretasi perkataan semantik diwakili dengan satu vektor baris konsep (artikel sepadan) berpasangan dengan pemberat yang menunjukkan darjah atau takat setiap konsep dikaitkan dengan perkataan. Kaedah ini boleh membina vektor konsep teks (ayat atau perenggan) sebagai sentroid vektor-perkataan dalam teks. Kajian lampau menunjukkan kaedah ESA ternyata dapat mengatasi kaedah terkini dalam menaksir hubungan semantik perkataan dan teks.

Taieb, Aouicha dan Hamadou (2013) mencadang kaedah gambaran semantik kategori untuk menggambar semantik bagi perkataan atas kategori Wikipedia selain daripada artikel dalam kaedah ESA. Setiap kategori Wikipedia diwakili sebagai vektor yang mengandungi terma dalam artikel, yang diklasifikasi kepada satu kategori. Terma bagi sebahagian artikel didefinisi sebagai set yang mengandungi frekuensi bagi setiap perkataan.

Kaedah Pengurangan Perwakilan Semantik (Saif, Ab Aziz & Omar, 2016) diperkenal bagi mengatasi kedimensian tinggi dalam kaedah analisis semantik tak tersirat. Lantaran itu, perwakilan semantik bagi perkataan diberi didefinisi sebagai konsep (ciri-ciri) yang mengandungi perkataan dalam maknanya tersendiri. Bagi pemberatan dan pengurangan ciri, tafsiran semantik bagi perkataan dibina sebagai vektor terhadap topik pendam daripada vektor perwakilan ESA yang asal. Bagi membentuk topik pendam, *Latent Dirichlet Allocation* disesuai dengan vektor ESA untuk mengeluar topik sebagai taburan kebarangkalian terhadap konsep selain daripada perkataan dalam model tradisional.

Kajian terbaharu Saif et al. (2017) mencadang satu Semantik Konsep Model (SCM) yang merupakan model kebarangkalian bagi menyepadu empat ciri semantik dalam Wikipedia. Model ini berdasarkan hipotesis yang pengabungan hubungan semantik yang berbeza membawa kepada model semantik jitu bagi meningkat keputusan dalam tugas pengukuran perkaitan semantik. Model SCM membina perwakilan semantik bagi konsep sebagai taburan

kebarangkalian terhadap konsep Wikipedia. Ia dibina di atas tanggapan semantik adalah konsep yang boleh diwakili dengan mengintegrasi ciri semantik seperti pautan templat, kategori, konsep penting dan topik. Konsep yang mempunyai kebarangkalian yang tinggi dalam model ini sering sinonim atau mempunyai hubungan semantik yang kuat dengan konsep asal.

PERBINCANGAN

Persamaan semantik dapat diklasifikasi ke dalam beberapa jenis pengukuran iaitu berasaskan laluan, berasaskan kandungan maklumat ke atas korpus, berasaskan kandungan maklumat ke atas taksonomi, berasaskan tindihan dan berasaskan ciri. Jadual 3 menunjukkan ringkasan dan kekurangan pengukuran persamaan semantik.

JADUAL 3 Ringkasan pengukuran persamaan semantik

Persamaan semantik	Kebaikan	Kekurangan
Berasaskan laluan	Ringkas dan mudah diimplementasi	Mengabai spesifikasi bagi konsep
Berasaskan kandungan maklumat ke atas korpus	Menggabung korpus dan WordNet untuk mendapat banyak persamaan semantik	Korpus memerlu anotasi manual yang memerlu banyak masa
Berasaskan kandungan maklumat ke atas taksonomi	Mengambil kira spesifikasi bagi konsep	Sensitif terhadap struktur taksonomi bagi sumber pengetahuan dan bergantung kepada kandungan maklumat ke atas dua konsep dan LCS sahaja
Berasaskan tindihan	Cekap dan memerlu glos serta konsep	Sensitif bagi memadan perkataan yang sinonim
Berasaskan ciri	Mudah diimplementasi dan fleksibel mengira persamaan semantik mengguna set teori	Andaian pemberatan yang malar pada semua unsur dalam perwakilan semantik konsep yang memiliki kaitan sama

Kertas ini, mengulas pengukuran semantik berasaskan pengetahuan yang berdasarkan sumber leksikal bagi menangkap bukti perkaitan semantik antara konsep. Pengukuran semantik dikelas kepada: (i) berasaskan laluan, (ii) kandungan maklumat, (iii) pengukuran berasaskan ciri, dan (iv) pengukuran berasaskan glos. Penyelidikan seperti yang dilakukan oleh (Ben Aouicha, Hadj Taieb & Ben Hamadou, 2016; Elavarasi, Akilandeswari & Menaga, 2014; Lastra-Díaz & García-Serrano, 2015; Zhang, Gentile & Ciravegna, 2013) hanya memfokus kepada pengukuran tradisional serta tidak membincang pengukuran taksonomi semantik berasaskan penggunaan struktur taksonomi sebagai ciri bagi perwakilan konsep. Justeru, kertas ini memfokus kepada pengukuran persamaan semantik berdasarkan struktur taksonomi.

Ukuran semantik berasaskan ciri (Batet, Sánchez & Valls, 2011; Ben Aouicha, Hadj Taieb & Hamadou, 2016; Gentleman 2005; Saif, Ab Aziz & Omar, 2014, 2017; Taieb, Aouicha & Hamadou, 2013, 2014) boleh didefinisi sebagai teknik yang bergantung kepada jumlah pengetahuan taksonomi yang dikongsi di antara dua konsep. Pengukuran tersebut memperkenal satu langkah berasaskan taksonomi baharu yang bergantung kepada struktur taksonomi dan teknik berasaskan tindihan bagi menentu-persamaan semantik di antara dua konsep. Ukuran kesamaan semantik adalah bersamaan nisbah antara jumlah bilangan bukan pengetahuan yang dikongsi dan jumlah dikongsi dan bukan pengetahuan yang dikongsi, dengan pengetahuan yang dikongsi di antara dua konsep didefinisi sebagai persilangan antara set super-konsep yang berasal dari taksonomi. Super-konsep bagi sesuatu konsep diekstrak sebagai nod laluan daripada nod yang sepadan kepada konsep seterusnya ke akar semantik taksonomi. Kelebihan utama ukuran berasaskan taksonomi ialah logik ‘is a’ hubungan antara konsep yang dikod oleh ahli leksikografi dalam taksonomi semantik. Selain daripada itu, pengukuran ini juga ditakrif oleh rumus yang mudah dan hanya memerlu taksonomi semantik

sebagai ukuran penilaian. Lantaran itu, ia boleh dilaksana dengan mudah, walaupun kos pengiraannya bergantung kepada saiz taksonomi semantik.

Isu utama yang dikenal pasti dalam pengukuran berasaskan ciri mengguna taksonomi ialah isu premis pemberatan seragam yang tidak jelas. Setiap elemen atau unsur dalam super-konsep dianggap mempunyai kaitan yang sama atau hubungan dalam pengiraan bagi keseluruhan panjang laluan tanpa mengira kedalaman sesuatu elemen atau unsur tersebut. Nod yang berbeza peringkat dalam semantik taksonomi mempunyai perbezaan nilai super-konsep. Nod yang berada pada peringkat yang rendah mempunyai lebih super-konsep berbanding nod yang berada pada peringkat tertinggi. Oleh kerana satu nod mempunyai lebih daripada satu laluan kepada akar (pelbagai hubungan hipernim), super-konsep diekstrak dari setiap laluan. Bagi spesifik nod, setiap elemen atau unsur dalam super-konsep yang terletak berdekatan dengan nod adalah lebih bermakna atau lebih semantik berbanding elemen atau unsur yang lain.

Perhubungan semantik di antara satu nod dan super-konsep adalah berbeza mengikut kedudukan nod dalam taksonomi. Nod yang berada pada kedudukan terendah adalah berkaitan atau semantik kepada super-konsep yang terdekat daripada nod yang cetek. Contoh, siratan hipernim langsung dari nod yang cetek sepadan kepada konsep *halogen* adalah konsep *group*, yang mengandungi perkataan *group, grouping* dan *clique*. Selain daripada itu, laluan daripada nod yang sepadan dengan konsep *Ox* kepada akar termasuk 15 nod adalah seperti berikut: *Ox→ Cattle→ Bovine→ Bovid→ Ruminant → Even-toed ungulate → Ungulate → Placental → Mammal → Vertebrate →Chordate →Animal →Organism →Living thing →Object →Entity →Root*. Pewaris terdekat konsep ini ialah *cattle, bovine, bovid* dan *ruminant*, yang mempunyai kata kunci berkaitan kepada konsep ini.

Bagi kaedah pengukuran Batet, tambahan kepada isu utama dalam pengukuran semantik berasaskan ciri terdapat dua masalah dalam pengukuran tersebut. Pertama, menghasil ketidakteraturan skor dalam pengukuran. Nilai persamaan semantik yang dijana mengguna ukuran ini mempunyai taburan yang tidak seragam jika mengguna konsep yang berbeza dalam sumber leksikal. Tambahan pula, nilai persamaan semantik ini juga tidak selari dengan nilai majoriti tahap pencapaian ukuran semantik yang lepas, nilai lingkungan ukuran semantik di selaras di antara sifar dan ketakterhinggaan. Isu lain yang timbul ialah apabila mengira persamaan semantik di antara dua perkataan yang mempunyai konsep yang serupa maka ini memberi nilai persamaan yang besar. Contoh, persamaan semantik antara *car* dan *automobile* ialah ketakterhinggaan kerana dua perkataan ini mempunyai konsep yang serupa.

KESIMPULAN

Bagi mencari penyelesaian terhadap permasalahan pembelajaran yang dikenal pasti dalam kaedah pengukuran persamaan semantik berasaskan ciri, kajian ini mencadang penggunaan parameter topologi seperti parameter tepi (*edge*), kedalaman (*depth*), keturunan (*descendant*) dan ketumpatan (*density*) bagi mengatasi andaian pemberatan yang malar bagi semua unsur dalam perwakilan semantik konsep yang mempunyai kaitan yang sama. Parameter struktur yang dicadang adalah penting bagi menghasil spesifikasi atau keluasan makna sesuatu konsep yang diguna secara meluas bagi mengemuka spesifikasi dalam ukuran persamaan semantik berasaskan laluan dan ukuran berasaskan kandungan maklumat. Selain daripada itu, mengeksploitasi model ruang vektor bagi mewakili semantik konsep selain daripada mengguna perwakilan semantik berasaskan set. Setiap unsur dalam ciri-ciri perwakilan tersebut dapat ditakrif dengan baik mengikut konsep yang sebetulnya dengan mengguna aplikasi teknik yang dicadang selain daripada prestasi pengukuran persamaan semantik dapat dijana dengan baik dan berkesan.

RUJUKAN

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M. & Soroa, A. 2009. A Study on Similarity and Relatedness Using Distributional and Wordnet-Based Approaches. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics, 19-27.

Azmi-Murad, M. & Martin, T. P. 2006. Sentence Extraction Using Asymmetric Word Similarity and Topic Similarity. In Abraham, A., De Baets, B., Köppen, M. & Nickolay, B. (eds) *Applied Soft Computing Technologies: The Challenge of Complexity*, 505-514. Berlin, Heidelberg: Springer.

Banerjee, S. & Pedersen, T. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. *In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. Acapulco, Mexico: Morgan Kaufmann Publishers, 805-810.

Batet, M., Sánchez, D. & Valls, A. 2011. An Ontology-Based Measure to Compute Semantic Similarity in Biomedicine. *Journal of Biomedical Informatics*, 44(1): 118-125.

Ben Aouicha, M., Hadj Taieb, M. A. & Ben Hamadou, A. 2016. Sisr: System for Integrating Semantic Relatedness and Similarity Measures. *Soft Computing*, (Online First) : 1-25.

Ben Aouicha, M., Hadj Taieb, M. A. & Hamadou, A. B. 2016. Taxonomy-Based Information Content and Wordnet-Wiktionary-Wikipedia Glosses for Semantic Relatedness. *Applied Intelligence*, (Online First : 1-37.

Budanitsky, A. & Hirst, G. 2006. Evaluating Wordnet-Based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1): 13-47.

Elavarasi, S. A., Akilandeswari, J. & Menaga, K. 2014. A Survey on Semantic Similarity Measure. *International Journal of Research in Advent Technology*, 2(3): 389-398.

Fellbaum, C. 1998. *Wordnet: An Electrical Lexical Database*. Cambridge, MA: The MIT Press.

Gabrilovich, E. & Markovitch, S. 2007. Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*. Hyderabad, India: Morgan Kaufmann Publishers, 1606-1611.

Gentleman, R. 2005. Visualizing and Distances Using Go. http://www.bioconductor.org/docs/vignettes.html [3rd May 2017].

Goodman, N. 1972. Seven Strictures on Similarity. In *Problems and Projects*, Indianapolis: Bobs-Merril.

Griffiths, T. L., Steyvers, M. & Tenenbaum, J. B. 2007. Topics in Semantic Representation. *Psychological Review*, 114(2):211-244.

Jiang, J. J. & Conrath, D. W. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *In Proceedings of International Conference Research on Computational Linguistics (ROCLING 97)*. Taipei; academia Sinica.

Jiang, Y., Zhang, X., Tang, Y. & Nie, R. 2015. Feature-Based Approaches to Semantic Similarity Assessment of Concepts Using Wikipedia. *Information Processing & Management*, 51(3):215-234.

Lastra-Díaz, J. J. & García-Serrano, A. 2015. A New Family of Information Content Models with an Experimental Survey on Wordnet. *Knowledge-Based Systems*, 89(C):509-526.

Leacock, C. & Chodorow, M. 1998. Combining Local Context and Wordnet Similarity for Word Sense Identification. *WordNet: An electronic lexical database*, 49(2): 265-283.

Lesk, M. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *Proceedings of the 5th annual International Conference on Systems Documentation*. Toronto, Canada: ACM, 24-26.

Li, Y., Bandar, Z. A. & Mclean, D. 2003. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):871-882.

Liberman, S. & Markovitch, S. 2009. Compact Hierarchical Explicit Semantic Representation. *Proceedings of the IJCAI 2009 Workshop on User-Contributed Knowledge and Artificial Intelligence: An Evolving Synergy (WikiAI09)*. Pasadena, CA: Morgan Kaufmann Publishers, 36-38.

Lin, D. 1998. An Information-Theoretic Definition of Similarity. *Proceedings of the 15th International Conference on Machine Learning (ICML)*. Madison, Wisconsin: Morgan Kaufmann Publishers, 296-304.

Liu, X.-Y., Zhou, Y.-M. & Zheng, R.-S. 2007. Measuring Semantic Similarity in Wordnet. *2007 International Conference on Machine Learning and Cybernetics*. 19-22 Aug, Hong Kong, China.

Medin, D. L., Goldstone, R. L. & Gentner, D. 1993. Respects for Similarity. *Psychological Review*, 100(2):254.

Meng, L., Gu, J. & Zhou, Z. 2012. A New Model of Information Content Based on Concept's Topology for Measuring Semantic Similarity in Wordnet. *International Journal of Grid & Distributed Computing*, 5(3):81-94.

Mihalcea, R., Corley, C. & Strapparava, C. 2006. Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity. *AAAI'06 Proceedings of the 21st National Conference on Artificial Intelligence*. Boston: AAAI Press, 775-780.

Milne, D. & Witten, I. H. 2008. Learning to Link with Wikipedia. *Proceedings of the 17th ACM conference on Information and knowledge management*. Napa Valley, California: ACM, 509-518.

Patwardhan, S., Banerjee, S. & Pedersen, T. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. *International Conference on Intelligent Text Processing and Computational Linguistics*. February 16 - 22, Mexico City, Mexico.

Petrakis, E. G., Varelas, G., Hliaoutakis, A. & Raftopoulou, P. 2006. X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies. *Journal of Digital Information Management*, 4(4):233-237.

Pirró, G. & Euzenat, J. 2010. A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. *In Proceedings of the 9th International Semantic Web Conference ISWC 2010*. Shanghai, China: Springer, 615-630.

Rada, R., Mili, H., Bicknell, E. & Blettner, M. 1989. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17-30.

Resnik, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Montreal: Morgan Kaufmann Publishers, 448–453.

Rodríguez, M. A. & Egenhofer, M. J. 2003. Determining Semantic Similarity among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2): 442-456.

Roget, P. M. 1911. *Roget's International Thesaurus, 1st Edition*. New York: Thomas Y. Crowell Co.

Saif, A., Ab Aziz, M. J. & Omar, N. 2014. Evaluating Knowledge-Based Semantic Measures on Arabic. *International Journal on Communications Antenna and Propagation*, 4(5): 80-194.

Saif, A., Ab Aziz, M. J. & Omar, N. 2016. Reducing Explicit Semantic Representation Vectors Using Latent Dirichlet Allocation. *Knowledge-Based Systems*, 100(May): 145-159.

Saif, A., Ab Aziz, M. J. & Omar, N. 2017. Mapping Arabic Wordnet Synsets to Wikipedia Articles Using Monolingual and Bilingual Features. *Natural Language Engineering*, 23(1): 53-91.

Saif, A., Omar, N., Aziz, M. J. A., Zainodin, U. Z. & Salim, N. 2017. Semantic Concept Model Using Wikipedia Semantic Features. *Journal of Information Science*, (Online First): Doi:10.1177/0165551517706231.

Sánchez, D. & Batet, M. 2013. A Semantic Similarity Method Based on Information Content Exploiting Multiple Ontologies. *Expert Systems with Applications*, 40(4): 1393-1399.

Sánchez, D., Batet, M. & Isern, D. 2011. Ontology-Based Information Content Computation. *Knowledge-Based Systems*, 24(2):297-303.

Sánchez, D., Batet, M., Isern, D. & Valls, A. 2012. Ontology-Based Semantic Similarity: A New Feature-Based Approach. *Expert Systems with Applications*, 39(9):7718-7728.

Seco, N., Veale, T. & Hayes, J. 2004. An Intrinsic Information Content Metric for Semantic Similarity in Wordnet. *16th European Conference on Artificial Intelligence, ECAI 2004, Including Prestigious Applicants of Intelligent Systems*. August 22-27, Valencia, Spain.

Strube, M. & Ponzetto, S. P. 2006. Wikirelate! Computing Semantic Relatedness Using Wikipedia. *In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*. Boston: AAAI Press, 1419-1424.

Taieb, H., Ben Aouicha, M., Tmar, M. & Hamadou, A. B. 2011. New Information Content Metric and Nominalization Relation for a New Wordnet-Based Method to Measure the Semantic Relatedness. *2011 IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS)*. 1-2 Sept, London.

Taieb, M. a. H., Aouicha, M. B. & Hamadou, A. B. 2013. Computing Semantic Relatedness Using Wikipedia Features. *Knowledge-Based Systems*, 50(Sept):260-278.

Taieb, M. a. H., Aouicha, M. B. & Hamadou, A. B. 2014. Ontology-Based Approach for Measuring Semantic Similarity. *Engineering Applications of Artificial Intelligence*, 36(Nov):238-261.

Turney, P. D. & Pantel, P. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141-188.

Wu, Z. & Palmer, M. 1994. Verbs Semantics and Lexical Selection. *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*. Las Cruces, New Mexico: Association for Computational Linguistics, 133-138.

Zesch, T. & Gurevych, I. 2010. Wisdom of Crowds Versus Wisdom of Linguists–Measuring the Semantic Relatedness of Words. *Natural Language Engineering*, 16(1):25-59.

Zhang, Z., Gentile, A. L. & Ciravegna, F. 2013. Recent Advances in Methods of Lexical Semantic Relatedness–a Survey. *Natural Language Engineering*, 19(04):411-479.

Zhou, Z., Wang, Y. & Gu, J. 2008. A New Model of Information Content for Semantic Similarity in Wordnet. *Second International Conference on Future Generation Communication and Networking Symposia (FGCNS'08)*. 13-15 Dec, Sanya, China.

Ummi Zakiah Zainodin[1], Nazlia Omar[2] & Abdul Gabbar Saif[3]
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor
e-mel:[1]ummizakiahzainodin@gmail.com
[2]nazlia@ukm.edu.my
[3]agmssaif@gmail.com

Received: 14 April 2017
Accepted: 14 June 2017